

07/17/00
07/17/00
U.S. POSTAL SERVICE
U.S. POSTAL SERVICE

07-19-00

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
UTILITY PATENT APPLICATION TRANSMITTAL
UNDER 37 CFR 1.53(b)

JC714 U.S. PTO A
09/617460
07/17/00

Address to: Assistant Commissioner for Patents Box Patent Application Washington, DC 20231	Attorney Docket No.	ARC9-2000-0116-US1
	Inventor(s)	KRAFT, Reiner
	Express Mail Label No.	EL537037696US
	Total Pages	61 (Not including references)

Title of Application:

**"SYSTEM AND METHOD FOR PREVENTING AUTOMATED
CRAWLER ACCESS TO WEB-BASED DATA SOURCES
USING A DYNAMIC DATA TRANSCODING SCHEME"**

Transmitted with the patent application are the following:

<input checked="" type="checkbox"/> 1 Page(s)	Transmittal form (and one copy)
<input checked="" type="checkbox"/> 40 Page(s)	Specification, claims, abstract
<input checked="" type="checkbox"/> 10 Page(s)	Formal Drawings
<input checked="" type="checkbox"/> 2 Page(s)	Declaration and Power of Attorney
<input checked="" type="checkbox"/> 1 Page(s)	Recordation of Assignment
<input checked="" type="checkbox"/> 2 Page(s)	Assignment of the Invention to International Business Machines Corporation
<input checked="" type="checkbox"/> 2 Page(s)	IDS and Form 1449A
<input checked="" type="checkbox"/> 2 References	Cited in Form 1449A
<input checked="" type="checkbox"/> 4 Page(s)	Return Receipt Postcard (MPEP 503)
<input checked="" type="checkbox"/> 1 Page(s)	Check No. 2069 to cover the filing fee
<input checked="" type="checkbox"/> 0 Page(s)	Certified copy of _____

This application is a: Continuation Divisional Continuation-in-Part of prior application Serial No. _____.

Fee Calculation

	Claims		Extra	Rate	Fees
Basic Fee					\$690.00
Total Claims	24	-20 =	4	× \$18.00	\$ 72.00
Independent Claims	3	- 3 =	0	× \$78.00	\$ 0.00
Multiple Dependent Claim				+	
				Assignm	\$ 40.00
				TOTAL	\$802.00

The Commissioner is hereby authorized to credit overpayments or charge fees required under 37 CFR 1.16 or 1.17 to Deposit Account No. 50-0219. Duplicate sheet attached.

EXPRESS MAIL CERTIFICATE

I hereby certify that the above paper/fee is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated below and is addressed to the Assistant Commissioner for Patents, Washington, DC 20231

Date of Deposit: 07/17/2000

Person mailing paper/fee: Samuel A. Kassatly

Signature _____

Respectfully submitted,
KRAFT, Reiner et al.



Samuel A. Kassatly
Reg. No. 32,247
Attorney for Applicants
Telephone: 408-323-5111
Samuel A. Kassatly Law Office
6819 Trinidad Drive
San Jose, CA 95120-2056

**SYSTEM AND METHOD FOR PREVENTING AUTOMATED
CRAWLER ACCESS TO WEB-BASED DATA SOURCES
USING A DYNAMIC DATA TRANSCODING SCHEME**

5

FIELD OF THE INVENTION

The present invention relates to the field of data processing, and particularly to web crawling technology. More specifically, this invention pertains to a computer software product for preventing crawlers from automatically accessing web-based data sources.

10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
115
120
125
130
135
140
145
150
155
160
165
170
175
180
185
190
195
200
205
210
215
220
225
230
235
240
245
250
255
260
265
270
275
280
285
290
295
300
305
310
315
320
325
330
335
340
345
350
355
360
365
370
375
380
385
390
395
400
405
410
415
420
425
430
435
440
445
450
455
460
465
470
475
480
485
490
495
500
505
510
515
520
525
530
535
540
545
550
555
560
565
570
575
580
585
590
595
600
605
610
615
620
625
630
635
640
645
650
655
660
665
670
675
680
685
690
695
700
705
710
715
720
725
730
735
740
745
750
755
760
765
770
775
780
785
790
795
800
805
810
815
820
825
830
835
840
845
850
855
860
865
870
875
880
885
890
895
900
905
910
915
920
925
930
935
940
945
950
955
960
965
970
975
980
985
990
995
1000
1005
1010
1015
1020
1025
1030
1035
1040
1045
1050
1055
1060
1065
1070
1075
1080
1085
1090
1095
1100
1105
1110
1115
1120
1125
1130
1135
1140
1145
1150
1155
1160
1165
1170
1175
1180
1185
1190
1195
1200
1205
1210
1215
1220
1225
1230
1235
1240
1245
1250
1255
1260
1265
1270
1275
1280
1285
1290
1295
1300
1305
1310
1315
1320
1325
1330
1335
1340
1345
1350
1355
1360
1365
1370
1375
1380
1385
1390
1395
1400
1405
1410
1415
1420
1425
1430
1435
1440
1445
1450
1455
1460
1465
1470
1475
1480
1485
1490
1495
1500
1505
1510
1515
1520
1525
1530
1535
1540
1545
1550
1555
1560
1565
1570
1575
1580
1585
1590
1595
1600
1605
1610
1615
1620
1625
1630
1635
1640
1645
1650
1655
1660
1665
1670
1675
1680
1685
1690
1695
1700
1705
1710
1715
1720
1725
1730
1735
1740
1745
1750
1755
1760
1765
1770
1775
1780
1785
1790
1795
1800
1805
1810
1815
1820
1825
1830
1835
1840
1845
1850
1855
1860
1865
1870
1875
1880
1885
1890
1895
1900
1905
1910
1915
1920
1925
1930
1935
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
2025
2030
2035
2040
2045
2050
2055
2060
2065
2070
2075
2080
2085
2090
2095
2100
2105
2110
2115
2120
2125
2130
2135
2140
2145
2150
2155
2160
2165
2170
2175
2180
2185
2190
2195
2200
2205
2210
2215
2220
2225
2230
2235
2240
2245
2250
2255
2260
2265
2270
2275
2280
2285
2290
2295
2300
2305
2310
2315
2320
2325
2330
2335
2340
2345
2350
2355
2360
2365
2370
2375
2380
2385
2390
2395
2400
2405
2410
2415
2420
2425
2430
2435
2440
2445
2450
2455
2460
2465
2470
2475
2480
2485
2490
2495
2500
2505
2510
2515
2520
2525
2530
2535
2540
2545
2550
2555
2560
2565
2570
2575
2580
2585
2590
2595
2600
2605
2610
2615
2620
2625
2630
2635
2640
2645
2650
2655
2660
2665
2670
2675
2680
2685
2690
2695
2700
2705
2710
2715
2720
2725
2730
2735
2740
2745
2750
2755
2760
2765
2770
2775
2780
2785
2790
2795
2800
2805
2810
2815
2820
2825
2830
2835
2840
2845
2850
2855
2860
2865
2870
2875
2880
2885
2890
2895
2900
2905
2910
2915
2920
2925
2930
2935
2940
2945
2950
2955
2960
2965
2970
2975
2980
2985
2990
2995
3000
3005
3010
3015
3020
3025
3030
3035
3040
3045
3050
3055
3060
3065
3070
3075
3080
3085
3090
3095
3100
3105
3110
3115
3120
3125
3130
3135
3140
3145
3150
3155
3160
3165
3170
3175
3180
3185
3190
3195
3200
3205
3210
3215
3220
3225
3230
3235
3240
3245
3250
3255
3260
3265
3270
3275
3280
3285
3290
3295
3300
3305
3310
3315
3320
3325
3330
3335
3340
3345
3350
3355
3360
3365
3370
3375
3380
3385
3390
3395
3400
3405
3410
3415
3420
3425
3430
3435
3440
3445
3450
3455
3460
3465
3470
3475
3480
3485
3490
3495
3500
3505
3510
3515
3520
3525
3530
3535
3540
3545
3550
3555
3560
3565
3570
3575
3580
3585
3590
3595
3600
3605
3610
3615
3620
3625
3630
3635
3640
3645
3650
3655
3660
3665
3670
3675
3680
3685
3690
3695
3700
3705
3710
3715
3720
3725
3730
3735
3740
3745
3750
3755
3760
3765
3770
3775
3780
3785
3790
3795
3800
3805
3810
3815
3820
3825
3830
3835
3840
3845
3850
3855
3860
3865
3870
3875
3880
3885
3890
3895
3900
3905
3910
3915
3920
3925
3930
3935
3940
3945
3950
3955
3960
3965
3970
3975
3980
3985
3990
3995
4000
4005
4010
4015
4020
4025
4030
4035
4040
4045
4050
4055
4060
4065
4070
4075
4080
4085
4090
4095
4100
4105
4110
4115
4120
4125
4130
4135
4140
4145
4150
4155
4160
4165
4170
4175
4180
4185
4190
4195
4200
4205
4210
4215
4220
4225
4230
4235
4240
4245
4250
4255
4260
4265
4270
4275
4280
4285
4290
4295
4300
4305
4310
4315
4320
4325
4330
4335
4340
4345
4350
4355
4360
4365
4370
4375
4380
4385
4390
4395
4400
4405
4410
4415
4420
4425
4430
4435
4440
4445
4450
4455
4460
4465
4470
4475
4480
4485
4490
4495
4500
4505
4510
4515
4520
4525
4530
4535
4540
4545
4550
4555
4560
4565
4570
4575
4580
4585
4590
4595
4600
4605
4610
4615
4620
4625
4630
4635
4640
4645
4650
4655
4660
4665
4670
4675
4680
4685
4690
4695
4700
4705
4710
4715
4720
4725
4730
4735
4740
4745
4750
4755
4760
4765
4770
4775
4780
4785
4790
4795
4800
4805
4810
4815
4820
4825
4830
4835
4840
4845
4850
4855
4860
4865
4870
4875
4880
4885
4890
4895
4900
4905
4910
4915
4920
4925
4930
4935
4940
4945
4950
4955
4960
4965
4970
4975
4980
4985
4990
4995
5000
5005
5010
5015
5020
5025
5030
5035
5040
5045
5050
5055
5060
5065
5070
5075
5080
5085
5090
5095
5100
5105
5110
5115
5120
5125
5130
5135
5140
5145
5150
5155
5160
5165
5170
5175
5180
5185
5190
5195
5200
5205
5210
5215
5220
5225
5230
5235
5240
5245
5250
5255
5260
5265
5270
5275
5280
5285
5290
5295
5300
5305
5310
5315
5320
5325
5330
5335
5340
5345
5350
5355
5360
5365
5370
5375
5380
5385
5390
5395
5400
5405
5410
5415
5420
5425
5430
5435
5440
5445
5450
5455
5460
5465
5470
5475
5480
5485
5490
5495
5500
5505
5510
5515
5520
5525
5530
5535
5540
5545
5550
5555
5560
5565
5570
5575
5580
5585
5590
5595
5600
5605
5610
5615
5620
5625
5630
5635
5640
5645
5650
5655
5660
5665
5670
5675
5680
5685
5690
5695
5700
5705
5710
5715
5720
5725
5730
5735
5740
5745
5750
5755
5760
5765
5770
5775
5780
5785
5790
5795
5800
5805
5810
5815
5820
5825
5830
5835
5840
5845
5850
5855
5860
5865
5870
5875
5880
5885
5890
5895
5900
5905
5910
5915
5920
5925
5930
5935
5940
5945
5950
5955
5960
5965
5970
5975
5980
5985
5990
5995
6000
6005
6010
6015
6020
6025
6030
6035
6040
6045
6050
6055
6060
6065
6070
6075
6080
6085
6090
6095
6100
6105
6110
6115
6120
6125
6130
6135
6140
6145
6150
6155
6160
6165
6170
6175
6180
6185
6190
6195
6200
6205
6210
6215
6220
6225
6230
6235
6240
6245
6250
6255
6260
6265
6270
6275
6280
6285
6290
6295
6300
6305
6310
6315
6320
6325
6330
6335
6340
6345
6350
6355
6360
6365
6370
6375
6380
6385
6390
6395
6400
6405
6410
6415
6420
6425
6430
6435
6440
6445
6450
6455
6460
6465
6470
6475
6480
6485
6490
6495
6500
6505
6510
6515
6520
6525
6530
6535
6540
6545
6550
6555
6560
6565
6570
6575
6580
6585
6590
6595
6600
6605
6610
6615
6620
6625
6630
6635
6640
6645
6650
6655
6660
6665
6670
6675
6680
6685
6690
6695
6700
6705
6710
6715
6720
6725
6730
6735
6740
6745
6750
6755
6760
6765
6770
6775
6780
6785
6790
6795
6800
6805
6810
6815
6820
6825
6830
6835
6840
6845
6850
6855
6860
6865
6870
6875
6880
6885
6890
6895
6900
6905
6910
6915
6920
6925
6930
6935
6940
6945
6950
6955
6960
6965
6970
6975
6980
6985
6990
6995
7000
7005
7010
7015
7020
7025
7030
7035
7040
7045
7050
7055
7060
7065
7070
7075
7080
7085
7090
7095
7100
7105
7110
7115
7120
7125
7130
7135
7140
7145
7150
7155
7160
7165
7170
7175
7180
7185
7190
7195
7200
7205
7210
7215
7220
7225
7230
7235
7240
7245
7250
7255
7260
7265
7270
7275
7280
7285
7290
7295
7300
7305
7310
7315
7320
7325
7330
7335
7340
7345
7350
7355
7360
7365
7370
7375
7380
7385
7390
7395
7400
7405
7410
7415
7420
7425
7430
7435
7440
7445
7450
7455
7460
7465
7470
7475
7480
7485
7490
7495
7500
7505
7510
7515
7520
7525
7530
7535
7540
7545
7550
7555
7560
7565
7570
7575
7580
7585
7590
7595
7600
7605
7610
7615
7620
7625
7630
7635
7640
7645
7650
7655
7660
7665
7670
7675
7680
7685
7690
7695
7700
7705
7710
7715
7720
7725
7730
7735
7740
7745
7750
7755
7760
7765
7770
7775
7780
7785
7790
7795
7800
7805
7810
7815
7820
7825
7830
7835
7840
7845
7850
7855
7860
7865
7870
7875
7880
7885
7890
7895
7900
7905
7910
7915
7920
7925
7930
7935
7940
7945
7950
7955
7960
7965
7970
7975
7980
7985
7990
7995
8000
8005
8010
8015
8020
8025
8030
8035
8040
8045
8050
8055
8060
8065
8070
8075
8080
8085
8090
8095
8100
8105
8110
8115
8120
8125
8130
8135
8140
8145
8150
8155
8160
8165
8170
8175
8180
8185
8190
8195
8200
8205
8210
8215
8220
8225
8230
8235
8240
8245
8250
8255
8260
8265
8270
8275
8280
8285
8290
8295
8300
8305
8310
8315
8320
8325
8330
8335
8340
8345
8350
8355
8360
8365
8370
8375
8380
8385
8390
8395
8400
8405
8410
8415
8420
8425
8430
8435
8440
8445
8450
8455
8460
8465
8470
8475
8480
8485
8490
8495
8500
8505
8510
8515
8520
8525
8530
8535
8540
8545
8550
8555
8560
8565
8570
8575
8580
8585
8590
8595
8600
8605
8610
8615
8620
8625
8630
8635
8640
8645
8650
8655
8660
8665
8670
8675
8680
8685
8690
8695
8700
8705
8710
8715
8720
8725
8730
8735
8740
8745
8750
8755
8760
8765
8770
8775
8780
8785
8790
8795
8800
8805
8810
8815
8820
8825
8830
8835
8840
8845
8850
8855
8860
8865
8870
8875
8880
8885
8890
8895
8900
8905
8910
8915
8920
8925
8930
8935
8940
8945
8950
8955
8960
8965
8970
8975
8980
8985
8990
8995
9000
9005
9010
9015
9020
9025
9030
9035
9040
9045
9050
9055
9060
9065
9070
9075
9080
9085
9090
9095
9100
9105
9110
9115
9120
9125
9130
9135
9140
9145
9150
9155
9160
9165
9170
9175
9180
9185
9190
9195
9200
9205
9210
9215
9220
9225
9230
9235
9240
9245
9250
9255
9260
9265
9270
9275
9280
9285
9290
9295
9300
9305
9310
9315
9320
9325
9330
9335
9340
9345
9350
9355
9360
9365
9370
9375
9380
9385
9390
9395
9400
9405
9410
9415
9420
9425
9430
9435
9440
9445
9450
9455
9460
946

10

20

The authors of web pages provide information known as metadata, within the body of the hypertext markup language (HTML) document that defines the web pages. A computer software product known as a web crawler, systematically accesses web pages by sequentially following hypertext links from page to page. The crawler indexes the pages for use by the search engines using information about a web page as provided by its address or Universal Resource Locator (URL), metadata, and other criteria found within the page. The crawler is run periodically to update previously stored data and to append information about newly created web pages. The information compiled by the crawler is stored in a metadata repository or database. The search engines search this repository to identify matches for the user-defined search rather than attempt to find matches in real time.

A conventional crawling operation 200 will now be briefly explained in connection with FIG. 2. A typical web crawler starts at block 201 by performing two main operations in order to execute the crawling process, namely, the access - retrieval of a document (block 202) and then the analysis phase of the document, also called the summarization process (block 204). Today's web crawler might be able to access a dynamically generated document, that is a document generated through executable code (e.g. CGI using Perl, ASP, C, or C++) on the web server.

Oftentimes, web designers embed an executable client side software code in the dynamic documents, so that eventually the code will be replaced with content or generates content on the client side. Examples of this executable code include

computation results that are originated based on some user input, or specific text based on the client's web browser version used. More generally, dynamic documents rely on a web browser's capabilities to:

5 a) retrieve additional documents (block 206) as needed or required, such as frames,

in-line images, audio, video, applets, or equivalents;

10 b) execute client side script (block 208) and code, such as JavaScript® or

equivalents;

15 c) furnish a fault tolerant HTML filter to recognize various HTML standards and

interpret HTML markup errors; unscramble content that a web designer has

20 purposefully scrambled in order to thwart crawling and other programmatic analysis

methods, to produce a final HTML markup (block 210); and

d) integrate all the previously obtained results to render the document (block 212)

for presentation to a user (block 214).

15 A typical search engine has an interface with a search window where the user enters

an alphanumeric search expression or keywords. The search engine sifts through

available web sites for the user's search terms, and returns the search of results in the

form of HTML pages. Each search result includes a list of individual entries that have

been identified by the search engine as satisfying the user's search expression. Each

20 entry or "hit" may include a hyperlink that points to a Uniform Resource Locator (URL)

location or web page.

Current web technology is being increasingly used for publishing and delivering

mission-critical information to consumers, customers, suppliers, and other entities. The extreme ease with which web data can be formed and published has been instrumental to the success and rapid adoption of the Internet as a preferred communication platform.

5

As explained herein, crawlers have been developed to automatically retrieve data from various web sites. The data may be used internally (e.g. competitive analysis) or externally (e.g. news feed aggregation). Crawlers can pose concerns to companies that publish their products and services on their web sites, desiring to make the data available to customers, to the exclusion of third parties aiming at invading the companies' own published data to entice customers away from these companies. Price data in particular, constitute sensitive information and a primary source of contention, since these data change frequently and can be the foundation of a price-leadership strategy.

15

20

There is currently no adequate mechanism by which the content of web pages can be protected from invading crawlers, without impacting the rendering of the web content to legitimate customers. This problem is further exacerbated by the difficulty in detecting crawlers and discriminating between crawler and web browser requests. The need for such a mechanism and corresponding process has heretofore remained unsatisfied.

SUMMARY OF THE INVENTION

The present protection system and associated method satisfy this need by making it significantly difficult, if not impossible for a crawler to automatically access and / or extract the company's web-based data. The present invention achieves this goal without impacting the ability of an interactive user, such as a consumer, to access the company's web-based data and to conduct regular business transactions.

For example, a customer using a web browser to access the company's web-based data, particularly price data, would still be able to access and interpret the data, while an automatic crawler will not be able to automatically process the data, even if the crawler were successful in accessing the data. In one embodiment, the protection system of the present invention will not prevent the crawler from downloading data. Rather, the data will be rendered non-extractable or automatically non-processable by the crawler.

The foregoing and further features and advantages of the present invention are achieved by any one of the following six transcoding techniques (e.g. software programs, algorithms, schemes, or processes), or by various combinations of these techniques:

The first transcoding technique involves dynamically changing the structure of a web form. Certain crawlers extract data by searching a web page for specific HTML elements and use them as anchors. This crawling technique will cause the search to fail

or to produce invalid results if the structure of the page were modified every time a web server sends the page to a client.

The second transcoding technique involves dynamically changing the content of a web form. Certain crawlers extract data by searching a web page for a keyword or phrase. Then, using the location of this keyword or phrase as an anchor, the crawlers extract the content data. Keyword searching can be prevented by inserting characters or images between the letters of the keyword or phrase that are invisible to an interactive user.

The third transcoding technique involves dynamically changing the names of form variables. Certain crawlers submit queries by hard-coding the names of form variables in URLs or the body of a POST message. However, these crawlers will fail if the names of the form variables are changed every session. A session may be established when the user visits the web site for the first time in a given time interval, such as the first visit of the day. Alternatively, every access to the web site may be viewed as a separate session.

The web server adds the session ID (identification) to the web form, which can then be extracted and used to compute synthetic names for the original form variables. The session ID is passed to the web user (crawler or browser) and the client passes the session ID back when the form is submitted. This allows the transcoding proxy to recompute the original name from the synthetic name. Computing synthetic names

generates fake (random, or meaningless) names for form variables. It utilizes a key in the generation of the fake names. The key is a session identifier, representing a session between a client and the server. The session ID is embedded within the web form. Recomputing the original names includes restoring the original names from the fake names and the key. The key again, is the session ID, and passed from the web client when the form is submitted.

The fourth transcoding technique involves using dynamic images to display text data. Many crawlers exclusively search the textual content of web page. These crawlers ignore inline images, even when these images contain critical text elements. According to this protection technique, the transcoding proxy alters the HTML content of the web page by replacing text elements with their image counterparts on the fly, thereby preventing a crawler from detecting the text elements.

15 The fifth transcoding technique involves using an executable application such as
JavaScript® to dynamically alter form variable names and content. Many crawlers
exclusively search the static HTML code of the web page. According to this technique,
pages that contain forms whose variables are initialized or calculated dynamically by a
JavaScript® code cannot be effectively crawled because the JavaScript® code would
20 have to be executed in order for the correct values to appear in the form variables.

The sixth transcoding technique involves using an executable application such as JavaScript® to dynamically generate the web page. Many robots exclusively search the

static HTML code of the web page. According to this technique, a web server can alter the HTML content of the web page by replacing the static HTML code with JavaScript® code fragments that "print" the HTML content when executed. Alternatively, the HTML content of the web page can be encrypted by the server and decrypted by the 5 JavaScript® code.

BRIEF DESCRIPTION OF THE DRAWINGS

The various features of the present invention and the manner of attaining them will be described in greater detail with reference to the following description, claims, and drawings, wherein reference numerals are reused, where appropriate, to indicate a correspondence between the referenced items, and wherein:

FIG. 1 is a schematic illustration of an exemplary operating environment in which an adaptive advertising system of the present invention can be used;

15 FIG. 2 is a flow diagram of a conventional web crawling or web browsing process;

20 FIG. 3 is a high level architecture of the crawling protection system of the present invention (FIG. 1) showing a transcoding proxy in use in an Internet context;

FIG. 4 is a more detailed block diagram of the transcoding proxy of FIG. 3,

illustrating six alternative or layered techniques available for use by the crawling protection system of FIG. 3;

FIG. 5 is a block diagram of the crawling protection system of FIGS. 1 and 3, shown

using a first transcoding technique according to the present invention;

5

FIG. 6 is a block diagram of the crawling protection system of FIGS. 1 and 3, shown

using a second transcoding technique according to the present invention;

10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95

FIG. 7 is a block diagram of the crawling protection system of FIGS. 1 and 3, shown

using a third transcoding technique according to the present invention;

FIG. 8 is a block diagram of the crawling protection system of FIGS. 1 and 3, shown

using a fourth transcoding technique according to the present invention;

FIG. 9 is a block diagram of the crawling protection system of FIGS. 1 and 3, shown

using a fifth transcoding technique according to the present invention; and

FIG. 10 is a block diagram of the crawling protection system of FIGS. 1 and 3,

shown using a sixth transcoding technique according to the present invention.

20

DETAILED DESCRIPTION OF THE INVENTION

The following definitions and explanations provide background information pertaining to the technical field of the present invention, and are intended to facilitate the understanding of the present invention without limiting its scope:

5

Crawler: Also referred to as "robot" or "spider", is a program that automatically explores the World Wide Web by retrieving a document and recursively retrieving some or all the documents that are linked to it.

10 002720 02150

15

HTML (Hypertext Markup Language): A standard language for attaching presentation and linking attributes to informational content within documents. During a document authoring stage, HTML "tags" are embedded within the informational content of the document. When the web document (or "HTML document") is subsequently transmitted by a web server to a web browser, the tags are interpreted by the browser and used to parse and display the document. In addition to specifying how the web browser is to display the document, HTML tags can be used to create hyperlinks to other web documents.

20

Internet: A collection of interconnected public and private computer networks that are linked together with routers by a set of standards protocols to form a global, distributed network.

Search engine: A remotely accessible World Wide Web tool that allows users to conduct keyword searches for information on the Internet.

5 Server: A software program or a computer that responds to requests from a web

browser by returning ("serving") web documents.

10 URL (Uniform Resource Locator): A unique address that fully specifies the location of a content object on the Internet. The general format of a URL is protocol://server-address/path/filename.

15 Web browser: A software program that allows users to request and read hypertext documents. The browser gives some means of viewing the contents of web documents and of navigating from one document to another.

20 Web document or page: A collection of data available on the World Wide Web and identified by a URL. In the simplest, most common case, a web page is a file written in HTML and stored on a web server. It is possible for the server to generate pages dynamically in response to a request from the user. A web page can be in any format that the browser or a helper application can display. The format is transmitted as part of the headers of the response as a MIME type, e.g. "text/html", "image/gif". An HTML web page will typically refer to other web pages and Internet resources by including hypertext links.

5 Web Site: A database or other collection of inter-linked hypertext documents ("web documents" or "web pages") and associated data entities, which is accessible via a computer network, and which forms part of a larger, distributed informational system such as the WWW. In general, a web site corresponds to a particular Internet domain name, and includes the content of a particular organization. Other types of web sites may include, for example, a hypertext database of a corporate "intranet" (i.e., an internal network which uses standard Internet protocols), or a site of a hypertext system that uses document retrieval protocols other than those of the WWW.

10 World Wide Web (WWW): An Internet client - server hypertext distributed information retrieval system.

15 FIG. 1 portrays the overall environment in which a crawling protection system 10 according to the present invention may be used. The system 10 includes a software or computer program product which is typically embedded within, or installed, at least in part, on a host server 15. Alternatively, the system 10 can be saved on a suitable storage medium such as a diskette, a CD, a hard drive, or like devices. While the system 10 will be described in connection with the WWW, the system 10 can be used with a stand-alone database of documents that may have been derived from the WWW 20 and/or other sources.

25 The cloud-like communication network 20 is comprised of communication lines and switches connecting servers such as servers 25, 27, to gateways such as gateway 30.

The servers 25, 27 and the gateway 30 provide the communication access to the WWW Internet. Users, such as remote Internet users are represented by a variety of computers such as computers 35, 37, 39, and can query the host server 15 for the desired information.

5

The host server 15 is connected to the network 20 via a communications link such as a telephone, cable, or satellite link. The servers 25, 27 can be connected via high speed Internet network lines 44, 46 to other computers and gateways. The servers 25, 27 provide access to stored information such as hypertext or web documents indicated generally at 50, 55, and 60. The hypertext documents 50, 55, 60 most likely include embedded hypertext link to other locally stored pages, and hypertext links 70, 72, 74, 76 to other webs sites or documents 55, 60 that are stored by various web servers such as the server 27.

15

20

FIG. 3 illustrates the crawling protection system 10 in use in an Internet context. The system 10 is represented by its main component, the transcoding proxy 315. FIG. 3 further shows a crawler 305 whose operation will be neutralized by the system 10, and a user's browser 310 whose operation will be allowed to proceed unaffected. According to one embodiment, the system 10 permutes the data accessed by the crawler 305 to render them uninterpretable by the crawler 305, while enabling the browser 310 to make use of the data. According to another embodiment, the crawler 305 will not be authorized to access the data, while the browser 310 will be allowed full access. This

functional duality is carried out without the system 10 differentiating between the crawler 305 and the browser 310.

FIG. 4 is a more detailed block diagram of the transcoding proxy 315 of FIG. 3, illustrating six alternative or layered techniques 351 - 356 that are available for use by the crawling protection system 10. The transcoding proxy 315 is generally comprised of a web page manipulator 330, a technique choosing module 340, and one or more protection techniques 351-356.

With further reference to FIG. 5, the input to the web page manipulator 330 includes an original web form 400 which is generated by the server 15. The original web form 400 contains data from its data sources 320 that are collected and organized in response to a request from either the crawler 305 or the browser 310. Data Sources 320 can include sensitive information, such as price information.

The web page manipulator outputs a dynamically manipulated web form 410 that prevents the crawler 305 from automatically accessing the data embedded in the original web form 400 without impacting the access from the browser 310. The manipulated web form 410 is the form resulting from the dynamic data transcoding of the original web form 400 by the transcoding proxy 315. The transcoding proxy 315 is an intermediary server that dynamically manipulates the original web form 400, using any one or more of the data transcoding techniques 351 - 356 described herein.

The web page manipulator 330 calls upon a technique choosing module 340 to select one or more of the transcoding techniques 351 - 356. The technique choosing module 340 keeps track of the various transcoding techniques 351 - 356, and selects the most appropriate combination to be applied to the original web page 400. One method of choosing the transcoding technique is through random access. The versatility of these techniques 351 - 356 is important to thwarting undesirable access by crawlers 305. While each of these transcoding technique will be described independently, it should be abundantly clear that one or all of these techniques 351 - 356 may be applied either simultaneously or sequentially. It should also be clear that the transcoding process is not limited to these six techniques, and that alternative variations can also be used.

The following is a list of the six data transcoding techniques 351-356 that form part of the transcoding proxy 315, and that are used for the prevention (or minimization) of automated crawler access to the web-based data stored in the data repositories 320:

- Transcoding technique that changes the web page structure.
- Transcoding technique that changes the web page content.
- Transcoding technique that selectively changes web page variable names.
- Transcoding technique that selectively converts text to images in the web page.
- Transcoding technique that alters form values when executed.
- Transcoding technique that generates a substantial portion of, or the entire web page when executed.

Several supporting routines can be utilized by the data transcoding techniques 351 - 356. For example, one particular supporting routine that can be used with the change of form variable names transcoding technique 353 (FIG. 4) invokes the computation of synthetic names (360) from the original names, and the recomputation of the original names from the synthetic names. This routine will be explained later in more detail.

5

First Transcoding Technique 351

Referring now to FIG. 5, it shows the crawling protection system 10 using a first transcoding technique 451 of the present invention. According to this illustrative example, either or both the crawler 305 and the browser (or graphic user interface, GUI) 310 send queries to the web server 15. In response to these queries, the web server 15 retrieves the desired data from the data sources or repository 320.

10
15
20

The web server 15 prepares the original web form 400 in a predetermined language, such as HTML, XML, or another equivalent language. The original web form 400 contains the content data and the structure or format in which the data will be presented to the user, regardless of whether the user is the crawler 305 or the browser 310. Typically, but not necessarily, the data format will be a column, row, or table 405. The example illustrated in FIG. 5 shows a price column 405 containing three rows.

20

The transcoding proxy 315 applies a transformation algorithm, also referred to herein as a first transcoding technique 351 to the original web form 400, to generate a manipulated web form 410. The web server 15 sends the same manipulated web form

410 to both the crawler 305 and the browser 310. However, this same form 410 will be unusable by the crawler 305, but will remain readable and browsable by the browser 310.

5 To this end, the first transcoding technique 451 involves dynamically changing the structure of the web form 400. Certain crawlers 305 extract data by searching the web page for specific HTML elements and use them as anchors. For example, the crawler 305 can search for, and use the third table, fifth row, second column as anchors. This crawling technique will cause the search to fail or to produce invalid results since the transcoding technique modifies the page structure every time a web server sends the page to the user (i.e., crawler 305, browser 310). In one embodiment, existing tables can be nested further, empty tables, rows or columns can be inserted, and/or simple lists can be replaced by tables.

10-15 15 In the present example, the original web form 400 includes a column of sensitive data information, namely the price data column 405. The transcoding proxy 315 transforms the original web form 400 into the manipulated web form 410, using the transcoding technique 351. According to a specific implementation, the transcoding technique 351 modifies the structure of the original price data column 405, but not its content, by, for example, automatically inserting an invisible character or string of characters, such as a blank line or space (also referred to herein as "structure inserts" 20 420.

As used herein, "content" refers to the substantive data contained in the original web form 400, and embedded within the language used by the web server 15, e.g. HTML. The content will be rendered visible to a user by means of the browser 310. "Structure" refers to the instructions from the web server 15 to the browser 310, directing the latter 5 to layout the content data according to a specific format. In regular use, the structure itself is not usually visible to the user. The crawler 305 typically uses the structure of a web page, would it be the original web form 400 or the manipulated web form 410, to interpret (locate or parse) the content data.

10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95

The modification to the original web form 400 can be accomplished using, for example a syntax change to the language (i.e., HTML, XML, etc.) in which the original web form 400 is written. The resulting transcoded price data column is referenced by 405T, and includes the same price data column 405 but moved to a different position, i.e., relocated, so that the data column 405 is more difficult to be automatically located by the crawler 305. In essence, the price data will be "hidden" from the crawler 305.

20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95

This structure manipulation will not affect the browser's ability to render and display the manipulated web form 410 (or the transcoded price column 405T), as the content data remain unaffected by the transcoding process. However, such structure change will impede the crawler's ability to interpret the content data. It should be noted that the data in the price data column 405 will be sent to the crawler 305, but will be rendered unusable, or very difficult to extract.

According to one embodiment, the transcoding technique 351 permutes all the data in the original web form 400 as described earlier. According to another embodiment, the transcoding technique permutes only selected data structures, such as price data columns, but not the remaining data. To this end, the web server 15 would need to pre-identify the sensitive data to be protected. According to yet another embodiment, the transcoding proxy 315 adds a fixed structure insert, such as a single row 420. However, according to still another embodiment, the transcoding proxy 315 randomly changes among the various structure inserts available to it. For example, the transcoding proxy 315 inserts one blank row 420 in the first data column, and inserts two blank rows 420 in the second data column, and so forth.

The transcoding technique 351 can be implemented using various alternatives. For example, the transformation of the original web form 400 can be done randomly or systematically. The data can be moved vertically and/or horizontally. Table nesting (i.e., a table with in a table) can also be used to "hide" the data.

Second Transcoding Technique 352

Referring now to FIG. 6, it shows the crawling protection system 10 using a second transcoding technique 452 of the present invention. Similarly to the foregoing description of FIG. 5, either or both the crawler 305 and the browser 310 send queries to the web server 15. In response to these queries, the web server 15 retrieves the desired data from the data sources or repository 320.

The web server 15 prepares the original web form 400 that contains the retrieved or content data and the structure in which the data will be presented to the user. In this example, the data will be rendered in a data column, row, or table 405. The example of FIG. 6 shows a price column 405 containing three rows.

5

The transcoding proxy 315 applies a transformation algorithm, also referred to herein as a second transcoding technique 352 to the original web form 400, to generate the manipulated web form 410. The web server 15 sends the same manipulated web form 410 to both the crawler 305 and the browser 310. However, this same form 410 will be unusable by the crawler 305, but will remain readable and browsable by the browser 310.

To this end, the second transcoding technique 452 involves dynamically changing the content of the web form 400 but not its structure. Certain crawlers 305 extract data by searching a web page for a keyword or phrase, which in this example is "Price". Then, using the location of this keyword or phrase as an anchor, the crawlers 305 extract the content data of interest, such as price value.

This transcoding technique 352 inserts characters, images, a string of characters or images, or otherwise data (referred to herein as "content inserts") that are invisible or substantially imperceptible to the user, between the letters of the keyword or phrase to prevent automatic keyword searching by the crawler 305. For instance, a crawler search for the keyword "price" would fail if this term were replaced by the following string:

10
15
20

5

"Price", which effectively inserts a blank space between the letter "P" and the subsequent letter "r", so that the word spells "P rice" 430. The space inserted between the letter "P" and the letter "r" in "Price" is imperceptible by the user, and could be rendered more so with a proper choice of the font in which the web page is rendered to the user.

This transcoding technique 352 can modify the original web form 400 on the fly by adding one or more invisible content inserts at random locations in selected keywords of interest that could lead to sensitive content data if used as anchors. This will produce a slightly different visual effect every time the web page or the server 15 is accessed. Slight deviations in the visual appearance of the page would be acceptable particularly that users will not notice or even pay attention to these deviations. In addition, the visual appearance of a web site is bound to be slightly different on different browsers.

This content manipulation will not affect the browser's ability to render and display the manipulated web form 410 (or the transcoded price column 405T), as the structure remains unaffected by the transcoding process, and the content data are unaltered except for the invisible inserts. However, these content inserts will impede the crawler's ability to automatically extract the content data since it prevents the crawler 305 from locating the desired keywords, effectively "blanking" the content data associated with these keywords. It should be noted that the data in the price data column 405 will be sent to the crawler 305, but will be rendered unusable, or very difficult to extract.

According to one embodiment, the transcoding technique 352 permutes all the keywords in the original web form 400 using one or more different content inserts. According to another embodiment, the transcoding technique permutes only selected keywords. According to yet another embodiment, the transcoding proxy 315 inserts a fixed content insert, such as a single space. However, according to still another embodiment, the transcoding proxy 315 randomly selects among various content inserts available to it. For example, the transcoding proxy 315 inserts one space in a first keyword, and inserts an image in the second keyword, and so forth.

The transcoding technique 352 can be implemented using various alternatives. For example, the transformation of the original web form 400 can be done randomly or systematically.

Third Transcoding Technique 353

Referring now to FIG. 7, it shows the crawling protection system 10 using a third transcoding technique 453 of the present invention. Either or both the crawler 305 and the browser 310 start by accessing the web server 15 that returns a log in page or form 500 to the requesting crawler 305 or browser 310. The log in page 500 typically includes one or more variables, each of which is comprised of a variable name 505 and a variable value. Conventionally, the crawler 305 or the browser (e.g. user) 310 fills in the variable values and submits the completed log in page 500 back to the web server 15. The web server 15 validates the variable names and values in the log in page 500, and

based thereon, retrieves the desired data from the data sources or repository 320 and forwards the retrieved data to the crawler 305 and/or the browser 310.

The transcoding proxy 315 applies a transformation algorithm, also referred to
5 herein as a third transcoding technique 353 to one or more of the variable names 505, to dynamically modify the variable names 505. Certain crawlers 305 submit queries by hard-coding the names of form variables in URLs or the body of a POST message. This is a valid approach because web applications are typically written using static variable names. However, because the transcoding proxy 315 modifies the variable names at
10 every session, the crawlers 305 will fail to log in and thus to access the data in the data sources 320. A session may be established when a user (crawler 305 or browser 310) visits the web site for the first time in a given time interval. Alternatively, every access to the web site may be viewed as a separate session.

In operation, the web server 15 receives a request to access the web site from the
15 crawler 305 and/or the browser 310, and generates the original log in page 500 with the required variables. The example of FIG. 7 illustrates the variables in the original log in page 500 as comprising two variable names 505: "username" and "password". The original log in page 500 also includes placeholders for variable values to be filled in by
20 the user. It should be clear that additional variables can alternatively be used.

The transcoding proxy 315 acts upon, and transforms one or more variable names so that the transformed or synthetic variable names 505T are not automatically

recognized by the crawler 305 as the proper names for which the variable values need to be inserted in the transformed log in page 510. In effect, the transcoding proxy 315 “hides” the original variable names 505 from the crawler’s view and replaces them with synthetic variable names 505T. The inability of the crawler 305 to recognize the original variable names 505 will prevent it from filling in the variable values in the corresponding placeholders, causing the access to the web server 15 to fail prior to accessing the data in the data sources 320.

However, neither the original variable names 505 nor synthetic variable names 505T are required by the browser 310 to enter the variable values, since they are not displayed by the browser 310. Instead, the browser 310 displays labels that correspond to the original variable names 505 and the placeholders for the variable values. Consequently, the transcoding technique 353 will be transparent to the browser 310 and does not affect the log in process or the log in presentation to the browser 310. Exemplary labels that correspond to the original variable names “username” and “password” are: “User Name” and “Password”, respectively. It is these labels that get displayed to the user of the browser 310.

Once the browser user fills in the required variable values and submits the log in page to the browser 15, the latter expects to receive the original variable names 505 as well, and to be able to differentiate the particular user’s log in data among various other users. The transcoding proxy 315 addresses and solves these concerns.

To this end, the web server 15 adds a session ID to the web form that can then be extracted and used to compute the synthetic names 505T for the original variable names 505 (block 360 of FIG 4). It is important that the session ID be passed to the browser 310 and that the browser 310 pass it back to the server 15 when the log in form 5 is submitted. This allows the transcoding proxy 315 to re-compute the original name from the synthetic names 505T (block 365 of FIG 4). This transcoding technique 353 may require only one key, e.g. the session ID, and assumes that the encoding is symmetric. Alternative schemes can use two or more keys.

10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95

10 The transcoding proxy 315 converts the synthetic variable names 505T to the original variable names 505 using the following process. The transcoding proxy 315 includes an encryption/decryption module 520 that encrypts the original variable names 500 and attaches an encryption key to the encrypted (or transcoded) variable names 505T, in the log in data stream from the web server 15 to the crawler 305 and/or browser 310.

15 When the browser 310 enters the required variable values and submits the transcoded log in page 510 back to the web server 15, the transcoded log in page 510 is intercepted by the transcoding proxy 315 so that the synthetic variable names 505T 20 are decrypted by the encryption/decryption module 520, back to the original variable names 505 using the same encryption key.

Fourth Transcoding Technique 354

Referring now to FIG. 8, it shows the crawling protection system 10 using a fourth transcoding technique 454 of the present invention. Either or both the crawler 305 and the browser 310 send queries to the web server 15. In response to these queries, the 5 web server 15 retrieves the desired data from the data sources or repository 320.

The web server 15 prepares the original web form 400 that contains the retrieved or content data and the structure in which the data will be presented to the user. In this example, the data will be rendered in a data column, row, or table 405. The example of FIG. 7 shows a price column 405 containing three rows.

The transcoding proxy 315 applies a transformation algorithm, also referred to herein as a fourth transcoding technique 354 to the original web form 400, to generate the manipulated web form 410. The web server 15 sends the same manipulated web form 410 to both the crawler 305 and the browser 310. However, this same form 410 will be unusable by the crawler 305, but will remain readable and browsable by the browser 310.

To this end, the fourth transcoding technique 454 involves using dynamic images to display text data. Many crawlers exclusively search the textual content of web pages. These crawlers ignore inline images, even when these images contain critical text elements. According to this protection technique, the transcoding proxy 315 alters the HTML content of the original web form 400 by replacing text elements, i.e., the price

column 405, with their image counterparts on the fly. The resulting manipulated web form 410 is readable by the browser 310, but not by the crawler 305 due to the crawler's inability to detect and the text elements that have been converted into images or to another similar format not readily recognizable by the crawler 305.

5

Fifth Transcoding Technique 355

Referring to FIG. 9, it shows the crawling protection system 10 using a fifth transcoding technique 455 of the present invention. Either or both the crawler 305 and the browser 310 start by accessing the web server 15 that returns a log in page or form 500 to the requesting crawler 305 or browser 310. The log in page 500 typically includes variable labels 550 and corresponding variable names 505 (FIG. 7) and variable values. Conventionally, the crawler 305 or the browser (e.g. user) 310 fills in the variable values that correspond to the labels 550, and submits the completed log in page 500 back to the web server 15. The web server 15 validates the variable names and values in the log in page 500, and based thereon, retrieves the desired data from the data sources or repository 320 and forwards the retrieved data to the crawler 305 and/or the browser 310.

The transcoding proxy 315 applies a transformation algorithm, also referred to herein as a fifth transcoding technique 355 to dynamically alter the form variable values entered by the user (e.g. the crawler 305, or the browser 310). This transcoding technique 355 involves using an executable application such as JavaScript® to dynamically alter the form variable values. Many crawlers exclusively search the static

HTML code of the web page. According to the present transcoding technique 355, pages that contain forms whose variables are initialized or calculated dynamically by a JavaScript® code cannot be effectively crawled because the JavaScript® code would have to be executed by the crawler 305, in order for the correct values to appear in the 5 form variables. Certain crawlers might include hard-coded rules for finding variable initialization statements in the JavaScript® code and thereby determining the correct values for the form variables. For these latter crawlers, the third transcoding technique 353 described above, can be used to prevent the crawlers from functioning properly.

10 In operation, the web server 15 receives a request to access the web site from the crawler 305 and/or the browser 310, and generates the original log in page 500 with the required labels and variables. The example of FIG. 9 illustrates two labels 550 contained in the original log in page 500: "User Name" and "Password". The original log in page 500 also includes placeholders for variable values to be filled in by the user. It 15 should be clear that additional variable labels can alternatively be included in the original log in page 500.

20 The transcoding proxy 315 acts upon, and transforms one or more labels 550 so that the new transformed or synthetic labels 550T become visible to the browser 310. The original labels 550 though still contained in the transcoded log in page 560 are not visible to the user of the browser 310, but are still detectable by the crawler 305.

The crawler 305 enters the variable values corresponding to the original labels 550 and submits these values to the server 15. However, this submission is intercepted by the transcoding proxy 315, which, using a programming or executable application, such as JavaScript®, overwrites the values entered by the crawler 305 in the original variables. This occurs because the proxy is programmed to copy values from the synthetic variables 550T into the original ones 550. This process results in blank or incorrect values to be submitted to the web server 15, which, upon seeing blank or incorrect variable values, denies crawler access to its data sources 320.

The browser 310 on the other hand, enters the variable values that are assigned to the visible, transformed or synthetic labels 550T and submits the transcoded log in page 560 to the transcoding proxy 315. The programming application, i.e., JavaScript®, residing in the transcoding proxy 315, copies the variable values entered by the browser 310 and assigns them to the original labels 550. The transcoded log in page 560, with the correct variable values corresponding to the original labels 550 are then submitted to the web server 15, authorizing browser access to the data sources 320.

Sixth Transcoding Technique 356

Referring now to FIG. 10, it shows the crawling protection system 10 using a sixth transcoding technique 456 of the present invention. Either or both the crawler 305 and the browser 310 send queries to the web server 15. In response to these queries, the web server 15 retrieves the desired data from the data sources or repository 320.

The web server 15 prepares the original web form 400 that contains the retrieved or content data and the structure in which the data will be presented to the user. In this example, the data is shown rendered in a data column or table 570.

5 The transcoding proxy 315 applies a transformation algorithm, also referred to herein as a sixth transcoding technique 356 to the original web form 570, to generate the manipulated web form 570T. The web server 15 sends the same manipulated web form 410 to both the crawler 305 and the browser 310. However, this same form 410 will be unusable by the crawler 305, but will remain readable and browsable by the browser 310.

10 To this end, the sixth transcoding technique 454 involves using an executable application such as JavaScript® to dynamically generate the web page to be displayed by the browser 310. Many robots exclusively search the static HTML code of the web page. According to this technique, a web server can alter the HTML content of the web page by replacing the static HTML code with JavaScript® code fragments that, for example, "print" the HTML content when executed. Alternatively, the HTML content of the web page can be encrypted by the server and decrypted by the JavaScript® code.

15 The resulting manipulated web form 410 is readable by the browser 310, but not by the crawler 305 due to the crawler's inability to allow the programming application to be executed, while the browser 310 would allow the execution of the application.

In operation, the web server 15 receives a request to access the web site from the crawler 305 and/or the browser 310, and generates the original web page 600 with the requested data 570 retrieved from the data sources 320. The transcoding proxy 315 acts upon, and transforms the original web page 600 into a manipulated web page 610.

5

The transcoding proxy 315 converts the structure and content of the original web page 600 which is written in a specific language, such as HTML code, to an executable application, such as JavaScript®, which, when executed, renders the original HTML code. While the same manipulated web page 610 is sent to the crawler 305 and/or the browser 310, the crawler 305 is incapable of recognizing the manipulated web page 610 prior to the execution of the JavaScript® code. In effect, the transcoding proxy 315 replaces the static content and/or structure of the original web page 600 with a dynamic or executable content and/or structure.

10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95

In a simplest form, the JavaScript® code can instruct the browser 310 to “write” the HTML content of the original web page 600, thus rendering visible to the user of the browser 310. To this end, the HTML code is embedded in the JavaScript® code, and the execution instructions of the JavaScript® code will recreate the original HTML code.

20

It is to be understood that the specific embodiments of the invention that have been described are merely illustrative of certain application of the principle of the present invention. Numerous modifications may be made to the crawling protection techniques and associated methods described herein, without departing from the spirit and scope

of the present invention. Moreover, while the present invention is described for illustration purpose only in relation to the WWW, it should be clear that the invention is applicable as well to databases and other tables with indexed entries.

5

What is claimed is:

1. A system for preventing automated crawler access to data from a network-based data source, comprising:

a transcoding proxy for automatically permutating data retrieved from the data source, to render the data uninterpretable by the crawler, while allowing a browser to render data retrieved from the data source.

2. A system for preventing automated crawler access to data from a network-based data source, comprising:

a transcoding proxy for automatically preventing crawler access to the data source, while allowing browser access to the data source.

3. The system according to claim 1, wherein the transcoding proxy further automatically prevents crawler access to the data source, while allowing browser access to the data source.

4. The system according to claim 1, wherein the transcoding proxy utilizes a transcoding technique that dynamically changes a form structure.

5. The system according to claim 1, wherein the transcoding proxy utilizes a transcoding technique that dynamically changes a form content.

6. The system according to claim 2, wherein the data retrieved from the data source include a variable with a name; and wherein the transcoding proxy utilizes a transcoding technique that dynamically changes the name of the variable.

7. The system according to claim 1, wherein the data retrieved from the data source include a textual content; and wherein the transcoding proxy utilizes a transcoding technique that dynamically changes the textual content into non-textual content.

8. The system according to claim 1, wherein the data retrieved from the data source includes a form containing a variable with a value; wherein the transcoding proxy utilizes a transcoding technique that incorporates an executable application in the form; and wherein when the executable application is implemented, it dynamically alter the value of the variable.

9. The system according to claim 1, wherein the data retrieved from the data source is collected in a page; wherein the transcoding proxy utilizes a transcoding technique that incorporates an executable application in the page; and wherein when the executable application is implemented, it dynamically generates the page and renders the data.

10. A system for preventing automated crawler access to data from a network-based data source, comprising:

a transcoding proxy for selectively, automatically permutating data retrieved from the data source, to render the data uninterpretable by one or more crawlers, while allowing a browser to render data retrieved from the data source; and

wherein the transcoding proxy further selectively, automatically prevents one or more crawlers from accessing the data source, while allowing browser access to the data source.

11. The system according to claim 10, wherein the data retrieved from the data source contain any one or more of: a variable with a name, a textual content, or a value;

wherein the data retrieved from the data source are collected in one or more of: a form or a page; and

wherein the transcoding proxy uses one or more of the following transcoding techniques:

a first transcoding technique that dynamically changes a form structure;

a second transcoding technique that dynamically changes a form content;

a third transcoding technique that dynamically changes the name of the variable;

a fourth transcoding technique that dynamically changes the textual content into

a non-textual content;

a fifth transcoding technique that incorporates an executable application in the form, and wherein when the executable application is implemented, it dynamically alter the value of the for variable; and/or

a sixth transcoding technique that incorporates an executable application in the page, and wherein when the executable application is implemented, it dynamically generates the page and renders the data.

12. The system according to claim 4, wherein the transcoding technique changes the form structure by shifting the position of the data in the form.

13. The system according to claim 12, wherein the transcoding technique shifts the position of form data by inserting any one or more of: a character, a string of characters, a space, non-textual data, a blank row, a blank column, and/or an empty table.

14. The system according to claim 13, wherein the transcoding technique shifts the position of form data by nesting the data in a table within a table.

15. The system according to claim 5, wherein the transcoding technique changes the form content by adding one or more content inserts that are substantially imperceptible to a browser user, in order to render selected terms in the page content difficult to be searched automatically by the crawler.

16. The system according to claim 15, wherein the content inserts include any one or more of: a characters, a string of characters, non-textual data, and/or a string of images.

17. The system according to claim 16, wherein the transcoding technique calculates a synthetic name from the variable name, and replaces the variable name with the synthetic name, in order to render selected terms difficult to be searched automatically by the crawler.

18. The system according to claim 17, wherein the variable is contained in a data entry form;

wherein the data entry form further includes a session ID; and

wherein the transcoding technique recalculates the variable name from the synthetic name to enable the data to be rendered by the browser.

19. The system according to claim 7, wherein the transcoding technique replaces the textual content by a corresponding image, in order to render selected terms difficult to be searched automatically by the crawler.

20. The system according to claim 8, wherein the variable includes a label associated with the variable;

wherein the transcoding technique further transforms the label into a synthetic label; and

wherein the executable application recalculates the variable value in order to prevent a server from authenticating the crawler access to the data source.

21. The system according to claim 9, wherein the transcoding technique renders the data using an HTML code.

22. A method for preventing automated crawler access to data from a network-based data source, comprising:

automatically permutating data retrieved from the data source, to render the data uninterpretable by the crawler, while allowing a browser to render data retrieved from the data source.

23. The method according to claim 22, further including automatically preventing crawler access to the data source, while allowing browser access to the data source.

24. The method according to claim 23, wherein the data contain any one or more of: a variable with a name, a textual content, or a value;

wherein the data are collected in one or more of: a form or a page; and
wherein automatically permuting data and preventing crawler access include using one or more of the following transcoding techniques:

a first transcoding technique that dynamically changes a form structure;

a second transcoding technique that dynamically changes a form content;

a third transcoding technique that dynamically changes the name of the variable;

a fourth transcoding technique that dynamically changes the textual content into a non-textual content;

a fifth transcoding technique that incorporates an executable application in the form, and wherein when the executable application is implemented, it dynamically alter the value of the for variable; and/or

a sixth transcoding technique that incorporates an executable application in the page, and wherein when the executable application is implemented, it dynamically generates the page and renders the data.

**SYSTEM AND METHOD FOR PREVENTING AUTOMATED
CRAWLER ACCESS TO WEB-BASED DATA
SOURCES USING A DYNAMIC DATA TRANSCODING SCHEME**

5

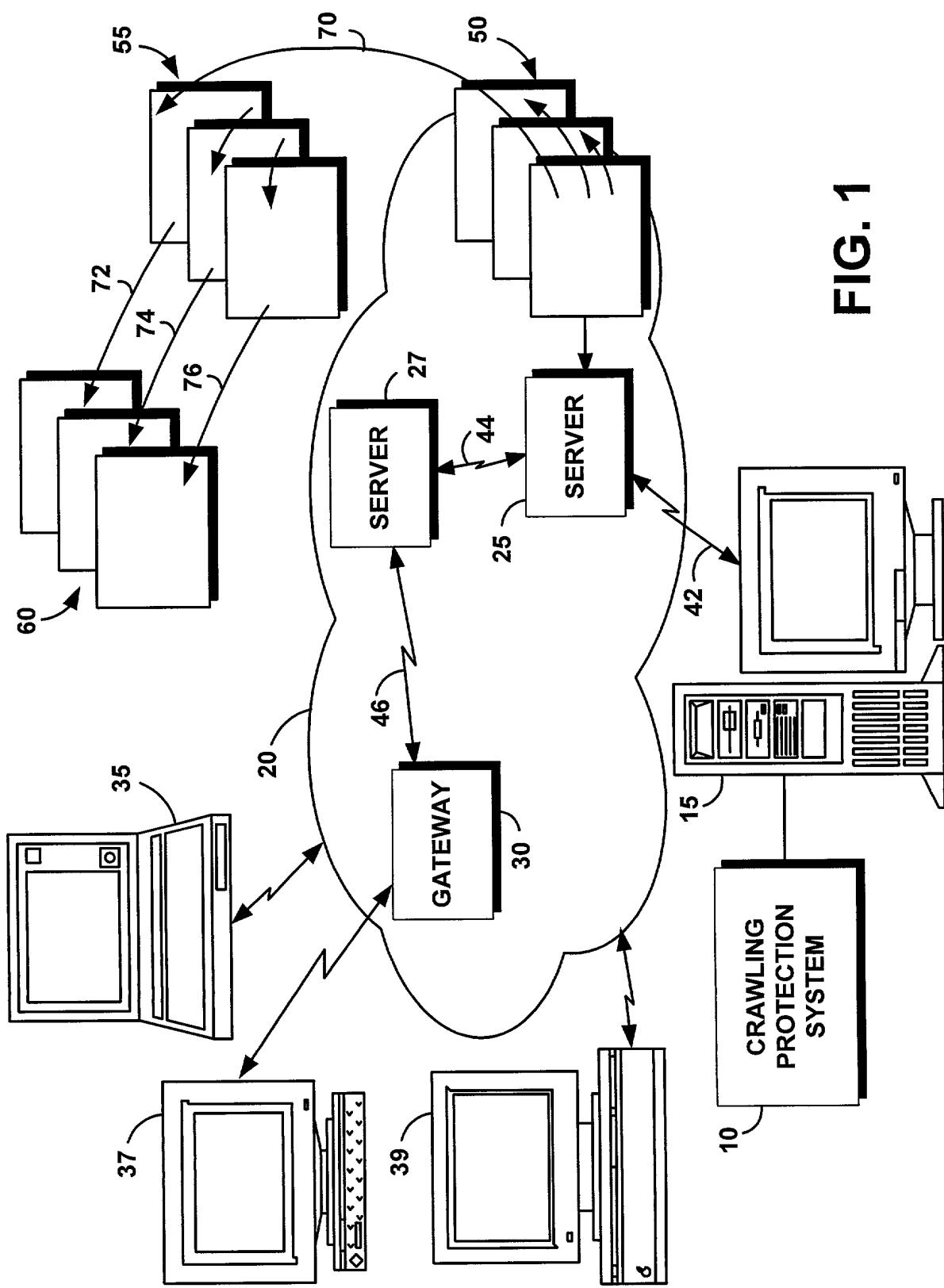
ABSTRACT OF THE INVENTION

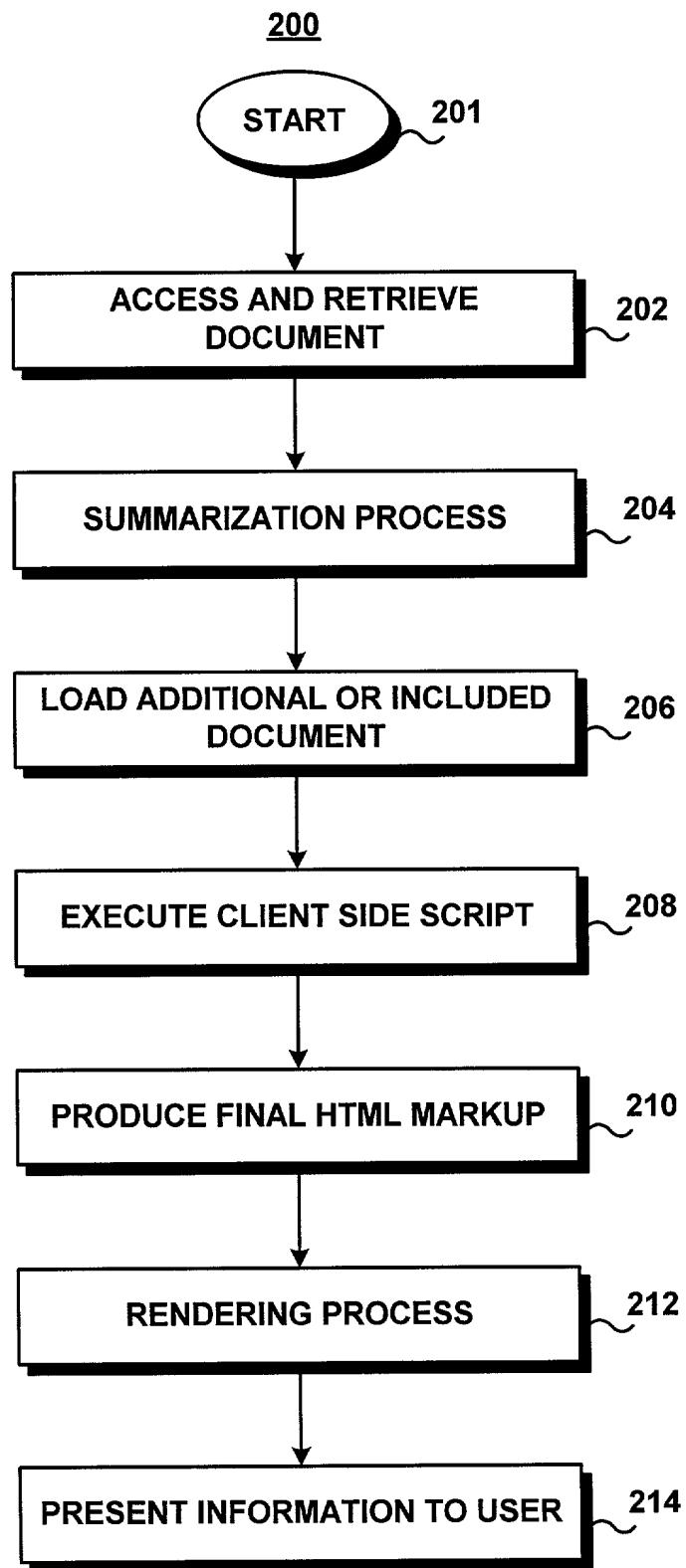
10
11
12
13
14
15
16
17
18
19
20

15

20

A protection system and associated method prevent the automatic crawler access to a company's web-based data, without impacting the ability of an interactive user, such as a consumer, to access the data and to conduct regular business transactions. In one embodiment, the protection system will not prevent the crawler from downloading data; rather, the data will be rendered non-extractable by the crawler. In another embodiment, the protection system will prevent crawler access to the data. To this end, the protection system uses any one or a combination of the following six transcoding techniques: Transcoding technique that changes the web page structure; transcoding technique that changes the web page content; transcoding technique that selectively changes web page variable names; transcoding technique that selectively converts text to images in the web page; transcoding technique that alters form values when executed; and/or transcoding technique that generates a substantial portion of, or the entire web page when executed.





PRIOR ART

FIG. 2

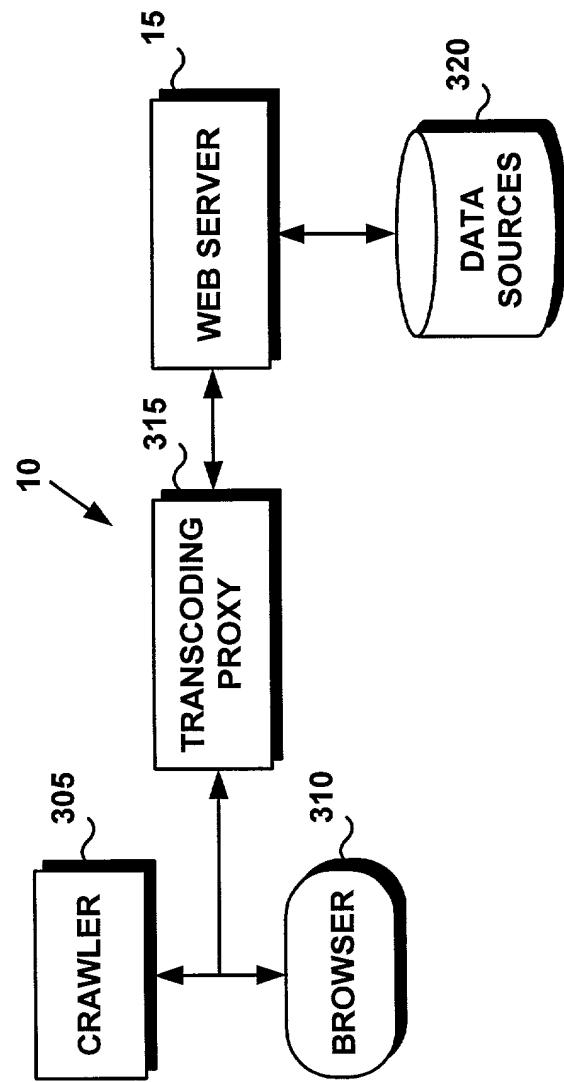


FIG. 3

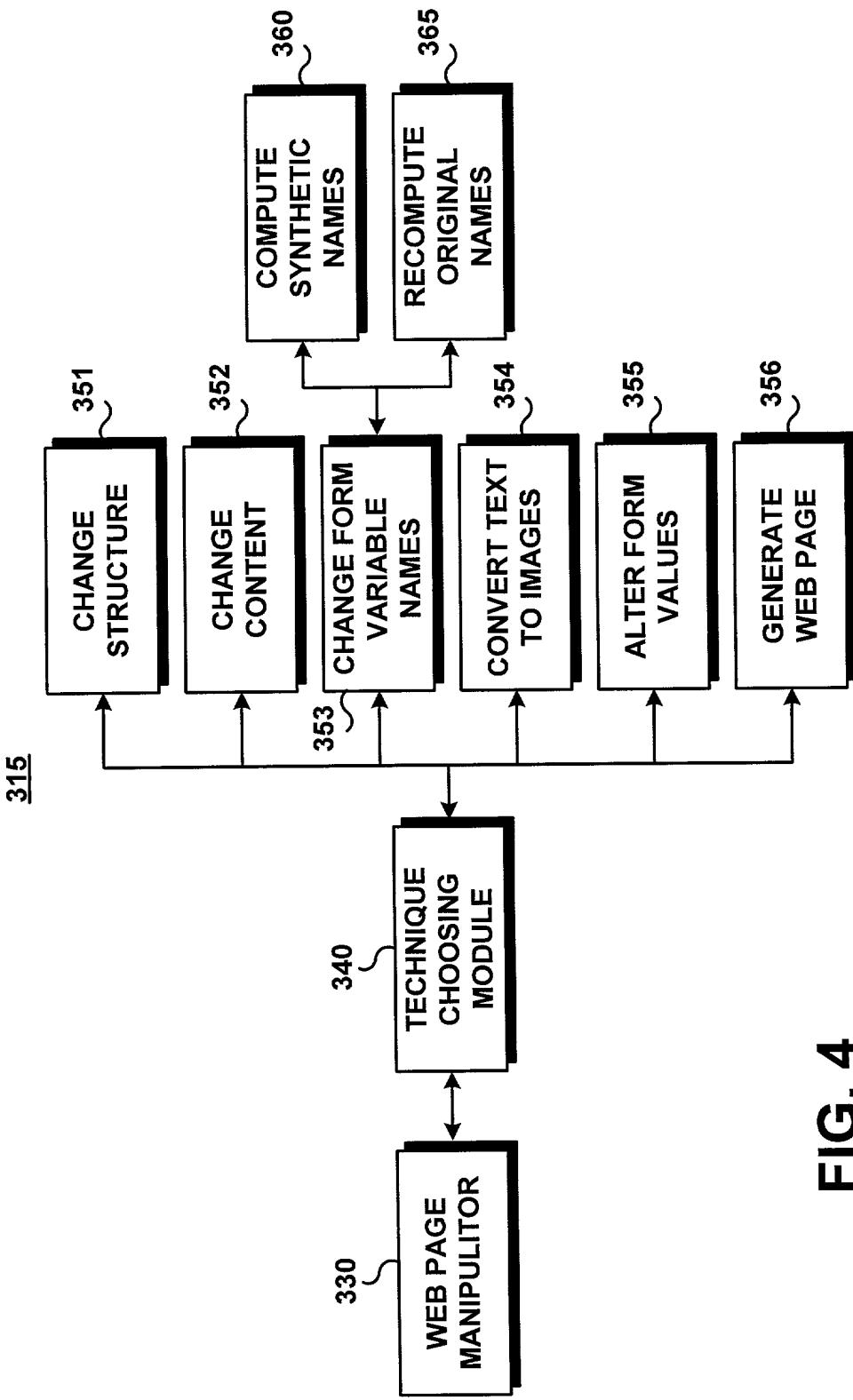


FIG. 4

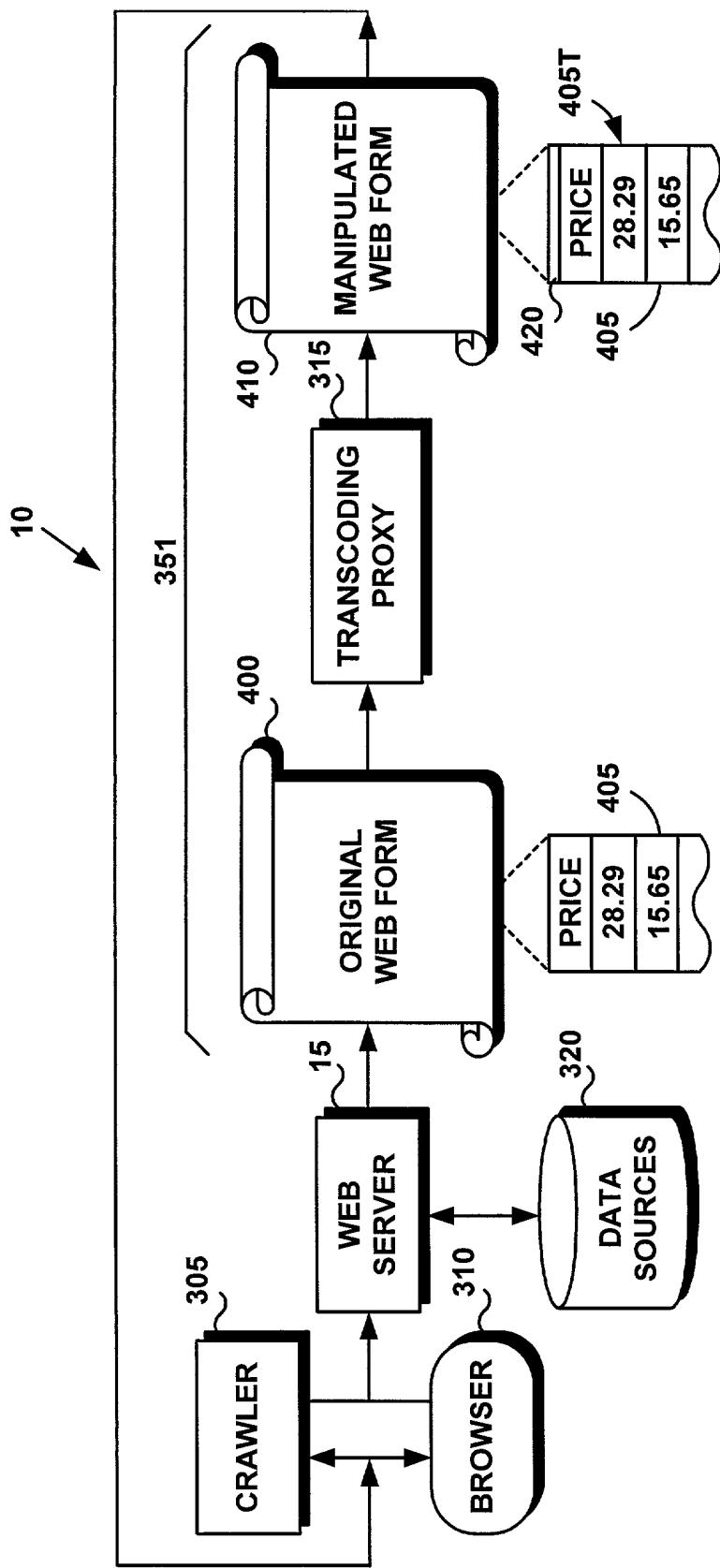


FIG. 5

FIG. 6

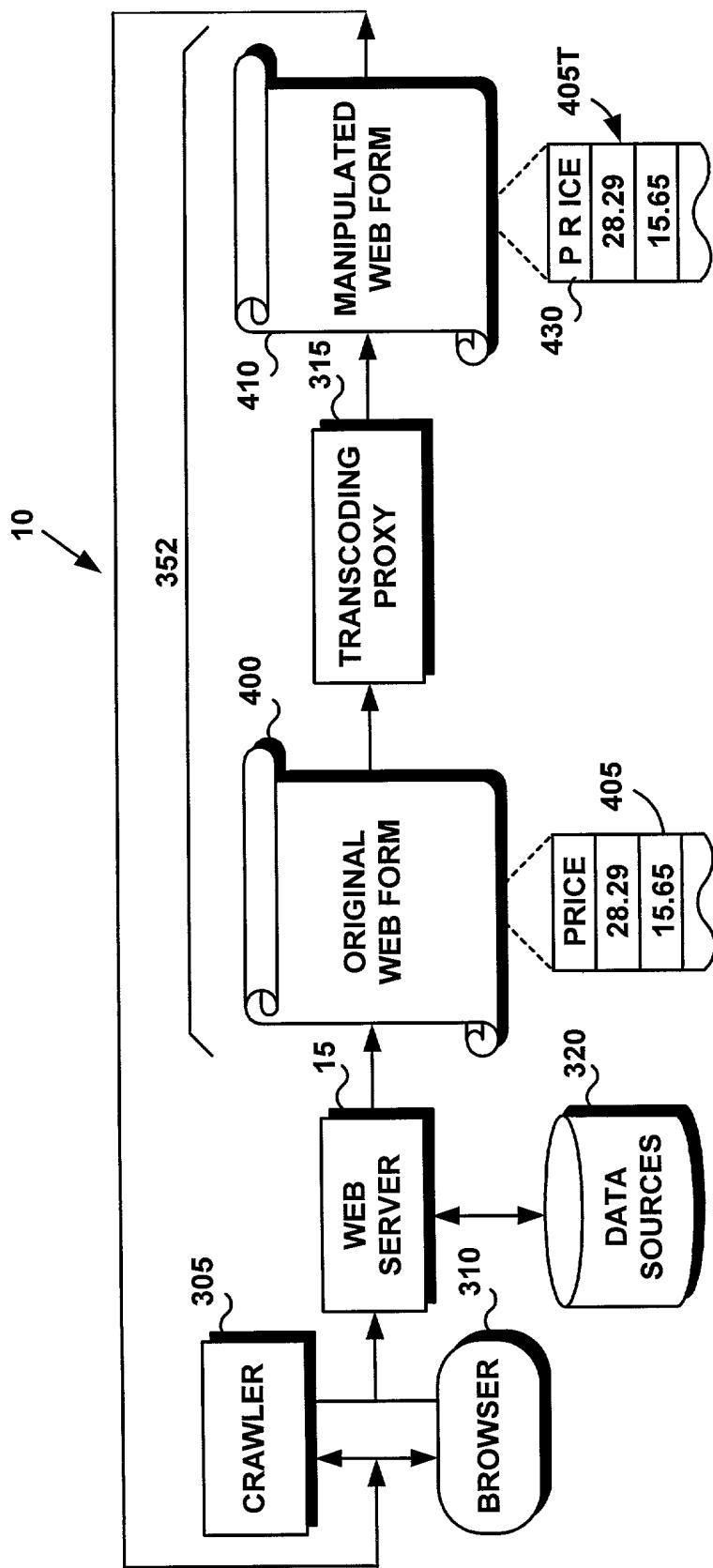


FIG. 7

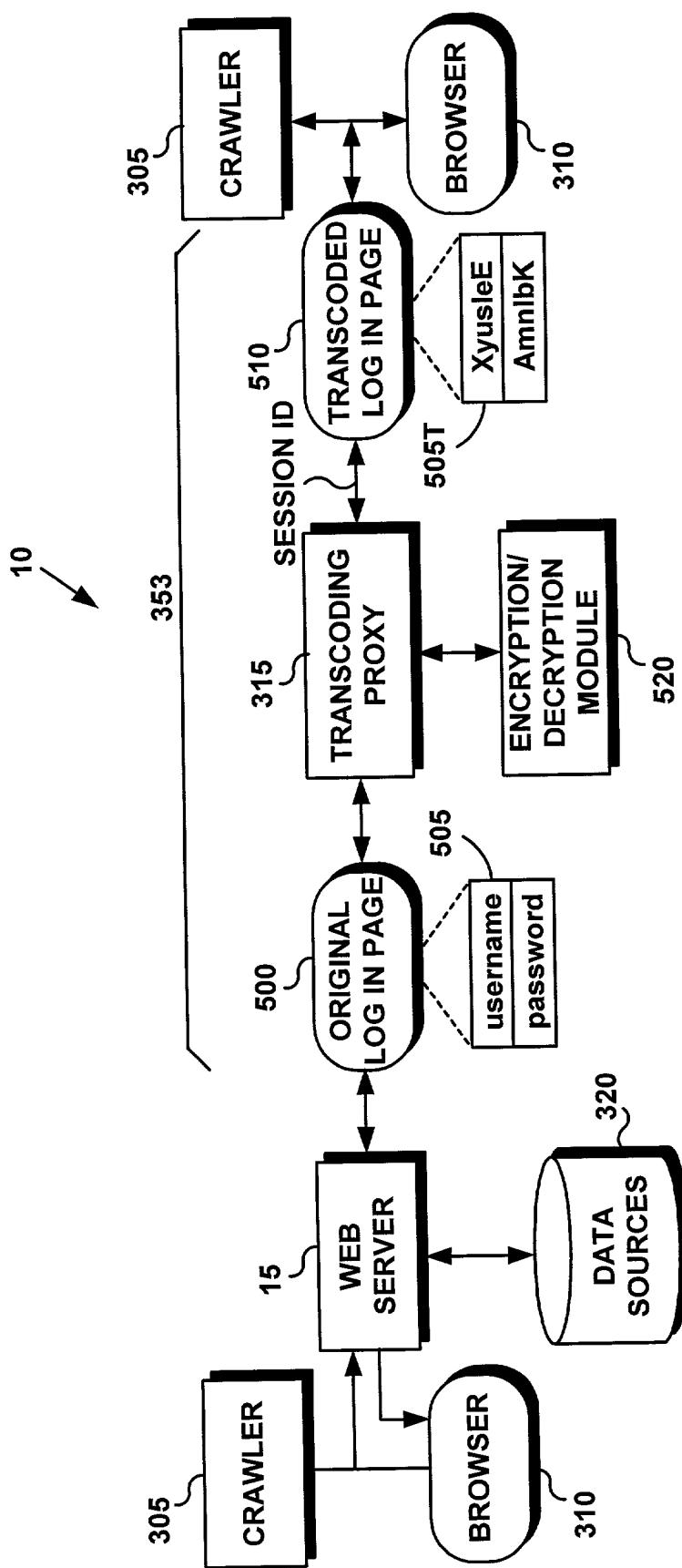


FIG. 8

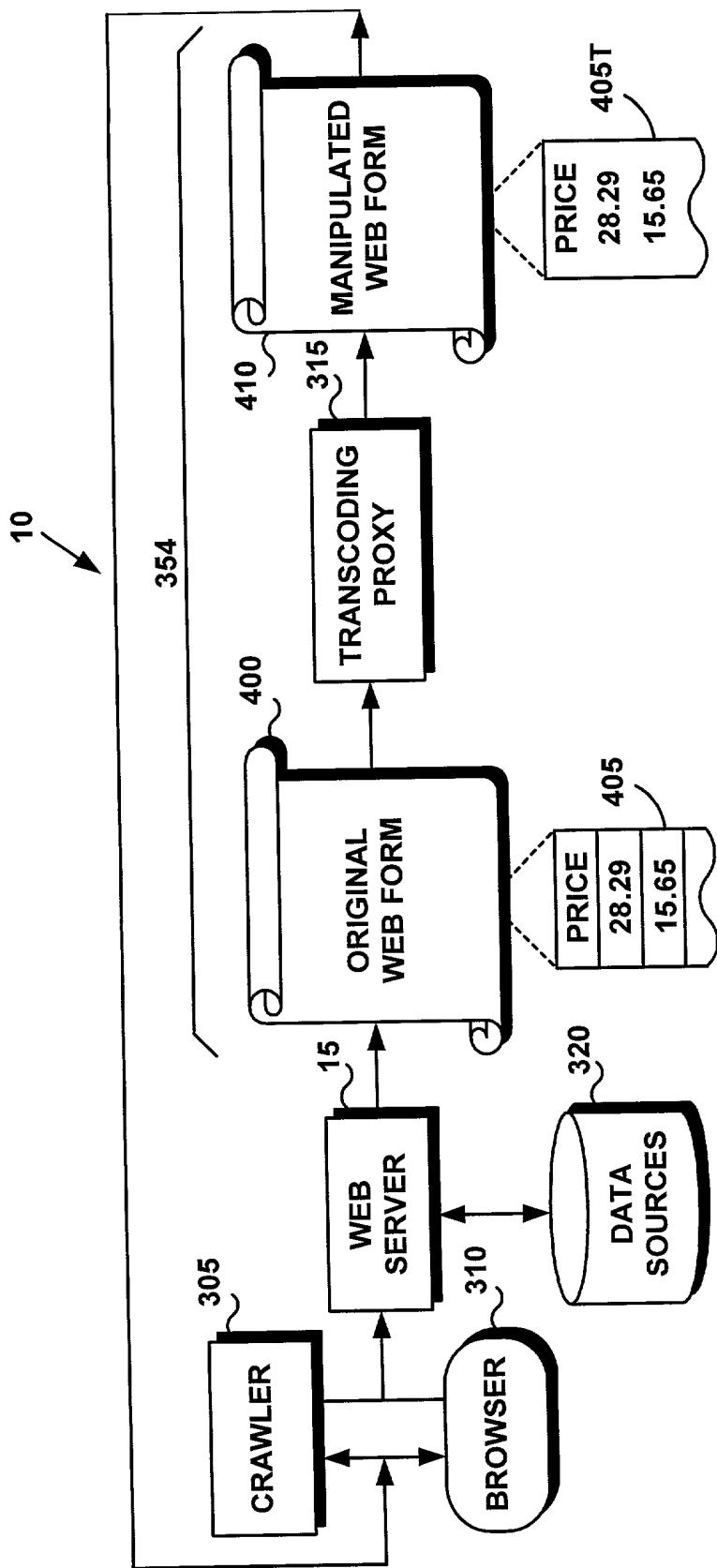


FIG. 9

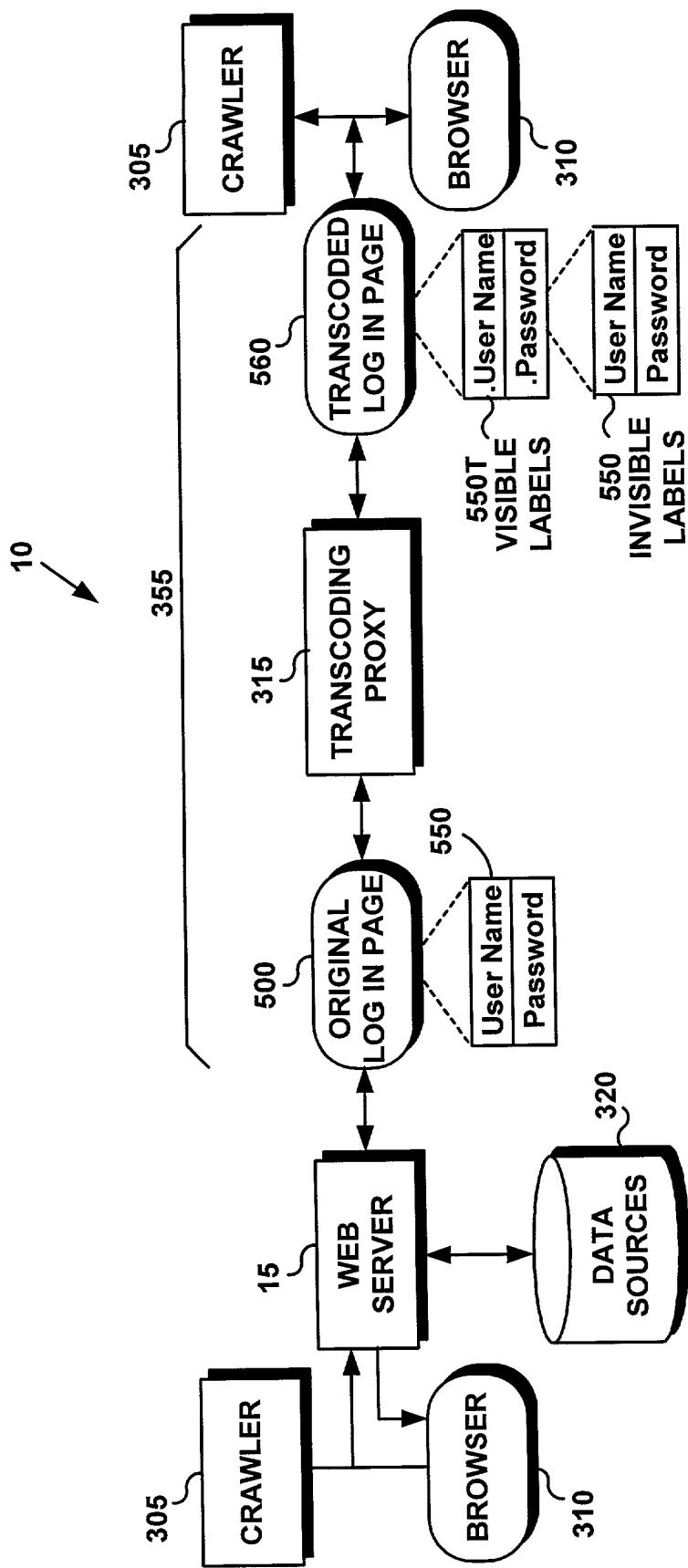
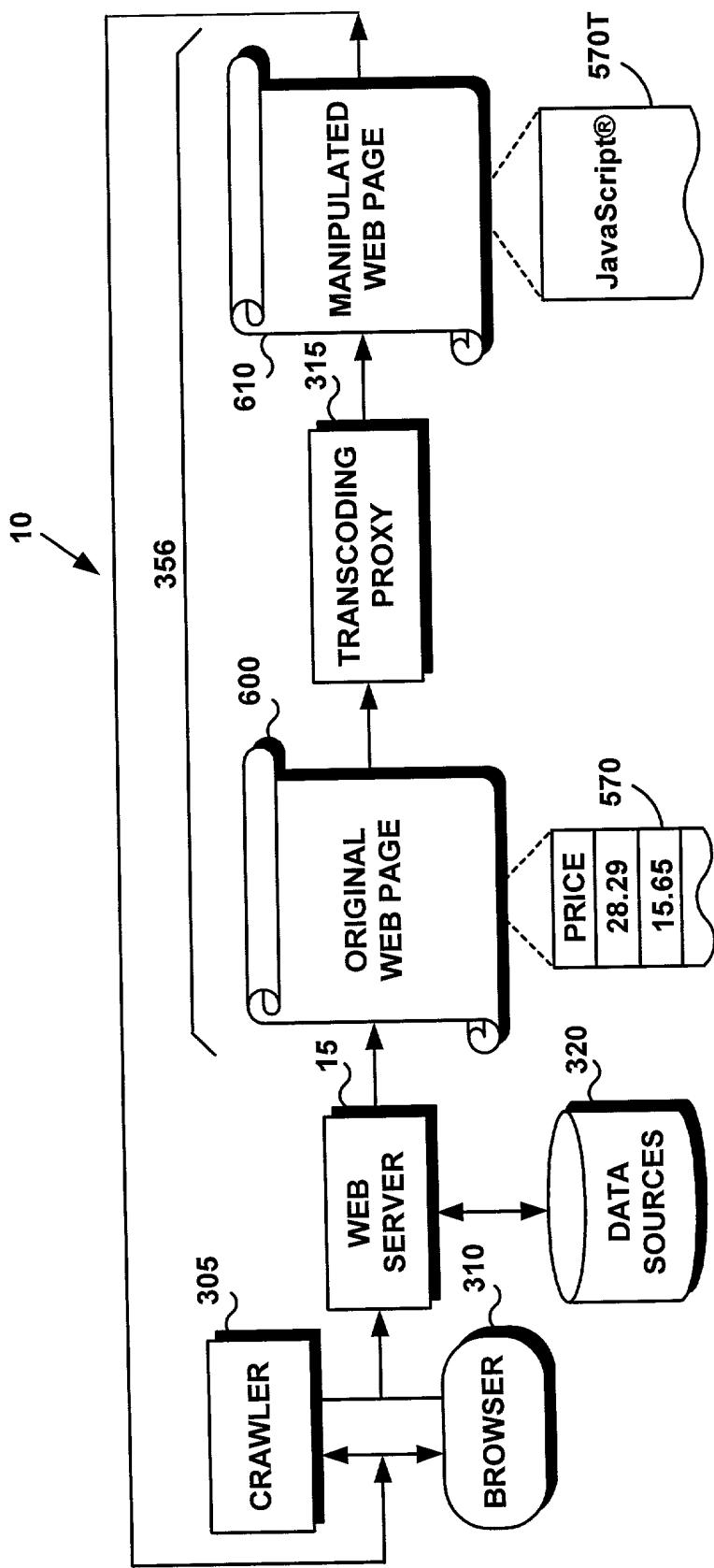


FIG. 10



DECLARATION AND POWER OF ATTORNEY FOR PATENT APPLICATION

Address all telephone calls to:
Samuel A. Kassatly
(408) 323-5111

Address all correspondence to:
Samuel A. Kassatly
6819 Trinidad Drive
San Jose, California 95120

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Full name of sole or first inventor: **KRAFT, Reiner**

Inventor's signature: 

Date:

7/17/2000

Residence: 9406 Wetsand Court
Gilroy, CA 95020

Citizenship: **Germany**

Post Office Address: **Same**

Full name of second inventor: **MYLLYMAKI, Jussi Petri**

Inventor's signature: 

Date:

7/17/2000

Residence: 395 Palm Ridge Lane
San Jose, CA 95123

Citizenship:

Finland

Post Office Address: **Same**

Full name of third inventor: **RUVOLO, Joann**

Inventor's signature: 

Date: 7/17/00

Residence: 6536 Leyland Park Drive
San Jose, CA 95120

Citizenship:

Finland

U.S.A. 7/17/00 Post Office Address: **Same**